A Graphical Check on Joint Gaussianity and a Modeled Variogram

Douglas M. Hawkins

Technical Report #547

April 1990

A Graphical Check on Joint Gaussianity and a Modeled Variogram

Douglas M Hawkins
Department of Applied Statistics
School of Statistics
University of Minnesota
St Paul MN

Abstract

Consider a regionalized variable Z(x). In many geostatistical contexts it is of interest to know whether Z(x) conforms to a multivariate Gaussian distribution, since if it does, then linear kriging is optimal and unbiased. The question is of even greater interest if lognormal kriging is planned, since Gaussianity of the log transformed grades is required even for lack of overall bias. A related question is whether the variogram is correctly modeled, since it is the variogram that provides the covariance matrix of any multigaussian distribution. This paper proposes an easy graphic check of the bivariate Gaussianity of pairs of data values. Its basis is that if a pair of data Z(x), Z(x+h) follows a bivariate Gaussian distribution with mean μ and a covariogram C(h), then the quadratic form defining their density follows an exponential distribution. The diagnostic proposed is a probability plot of the quadratic forms, where a straight line supports the model of Gaussianity with the assumed variogram. Departures from the model such as non-Gaussianity, outliers, and a wrong variogram produce characteristic non-linear behavior of the plot.

Keywords Probability plot, Gaussian models, Lognormal kriging, Optimality.

Introduction

Consider a regionalized variable (ReV) Z with finite first and second moments which is believed to follow a multivariate Gaussian distribution with mean μ and semivariogram $\gamma(h)$.

While the converse is not true, multivariate Gaussianity implies bivariate Gaussianity, and so a check on whether the data on the ReV conform to bivariate Gaussianity can give some support to, or can rule out, multivariate Gaussianity of the ReV. The bivariate Gaussian model for a pair of observations Z(x+h) and Z(x), along with the assumption of finite mean μ and variance σ^2 and specified semivariogram $\gamma(h)$ mean that the joint distribution of Z(x+h) and Z(x) can be written

$K \exp[-\frac{1}{2}W(x,h)]$ where the quadratic form of the density is

$$W(x,h) = \left[\frac{\{Z(x+h)-Z(x)\}^2+2\tau(h)\{Z(x+h)-\mu\}\{\{Z(x)-\mu\}}{\tau(h)\{2-\tau(h)\}\sigma^2}\right]$$

with $\tau(h) = \gamma(h)/\sigma^2$, and K is a normalizing constant whose value is not of concern here.

Under the bivariate Gaussian model, the distribution of the quadratic form W(x,h) is very simple - it is chi-squared with 2 degrees of freedom - ie its cumulative distribution function is

$$\Pr[W(x,h)<w] = 1 - e^{-\frac{1}{2}w}$$

Equivalently, transforming W(x,h) to $U(x,h) = 1-e^{-\frac{1}{2}W(x,h)}$, the U(x,h) follow a uniform distribution over the range (0,1).

This lends itself to a very simple graphic check on distribution - the probability-probability or P-P plot. Given a sample of values W(x,h), compute the corresponding U(x,h) values and plot their cumulative distribution function. If the W(x,h) values do follow the assumed distribution, then the plot will approximate the straight line from (0,0) to (1,1). This leads to our first proposal.

Proposal 1. For each pair of Z values Z(x+h), Z(x) in the data, compute the quadratic form W(x,h) transform to U(x,h). Divide the range (0,1) into (say) 100 equal width cells, and count the number U values in each cell. Plot the cumulative distribution function based on these values. If the plot 1 45 degree line from (0,0) to (1,1), this is evidence of fit of the model of bivariate Gaussianity wit assumed variogram. If it does not conform to the straight line, then this is proof that the model us valid. Further checks are then carried out to see whether the model fails because of non-Gaussian a mis-specified variogram.

There are several possible checks of the conformity of the plot to the 45 degree line. The first is the Kolmogorov test statistic - the maximum deviation between the empiric distribution function and the target straight line. Some caution however is needed in testing the value of the Kolmogorov statistic formally. In a sample of n values of the regionalized variable, there will be $N = \frac{1}{2}n(n-1)$ different U(x,h) values. However these do not constitute an independent sample of size N, and so even though the P-P plot is based on N values, it is not appropriate to test the Kolmogorov test statistic as if it were based on an independent sample of size N.

We are not aware of theoretical results on the true null distribution of these interdependent values, but a reasonable and conservative approach would be to use n, the number of genuinely independent values. The 95% point of the Kolmogorov statistic for a sample of size n is approximately $1.22/\sqrt{n}$, provided n is bigger than 20, as will generally be the case in geostatistical data sets.

The Kolmogorov statistic K is sensitive to the general shape of the distribution of the U. Some departures from the model manifest themselves in the frequency of very large W(x,h) values, and so K needs to be supplemented by other measures sensitive to tail behavior. One such check is on the single largest W(x,h) value W^* say. The probability that the largest of N independent W(x,h) values would attain or exceeds W^* can be bounded by the Bonferroni inequality:-

$$\Pr[W^* < w] \leq Ne^{-\frac{1}{2}w}.$$

Thus the statistic $Q = Ne^{-\frac{1}{2}W^*}$ provides a conservative test of whether the largest W(x,h) is significantly outlying. For example, if Q<).05, then this indicates at a conservative 5% significance level that the largest W(x,h) value does not conform to the model.

A broader check on, not just the single largest W(x,h), but the preponderance of large values is obtained by letting L be the proportion of the data in the (0.99,1.00) cell, which should be close to 0.01. If the N U(x,h) values were independent, then the number of values falling in this cell would be binomially distributed, and so have mean 0.01 and variance 0.0099/N. However the U(x,h) values are not independent, and so this binomial distribution does not apply. Once again however, it is reasonable to approximate the variance of the proportion conservatively, by basing it on the number of distinct data values, as 0.0099/n. This gives the operational check of finding the test statistic

$V = \sqrt{n}(10L-0.1)$
and assessing it against the standard normal distribution.

It should be mentioned that while bivariate Gaussianity implies the chi-squared distribution for W, the converse is not necessarily true. It is possible for W to follow a chi-squared distribution while the Z values are not bivariate Gaussian, but it is not particularly common. See for example Hawkins (1981). Cox and Small (1978) suggest a way of checking for the possibility that the data are not multivariate Gaussian even though quadratic form is chi-squared.

Effect of departures from Gaussian.

We now consider various common departures from model and their effects on the three diagnostics K (the Kolmogorov statistic), Q (the significance of the largest W(x,h)), and V (the frequency of the top centile).

(1) Heavy-tailed non-normality, right variogram. Suppose the distribution is stationary with a non-normal heavy-tailed distribution but with the mean, variance and variogram supposed - $E[Z(x)] = \mu$, $E[Z(x)-\mu]^2 = \sigma^2$ and $E[Z(x+h)-Z(x)]^2 = 2\gamma(h)$. The effect of the non-normality will be to create an excess of W(x,h) values at the two ends of the scale and a deficiency in the middle. This will have two effects - (i) to increase the number of U(x,h) values in the top few cells, and(ii) to make the main part of the plot a convex curve lying below the target line. This will tend to give significantly large values of V, and K, and a small value for Q.

The mirror image, but uncommon departure of a light-tailed distribution will give a concave plot lying above the line, with a shortage of large U(x,h) values. This will manifest itself in a large value of K, a large negative value for V and a large (and therefore insignificant) value for Q.

(2) Outlying values The second possible departures is the occurrence of a small number of outlying values. Both operationally and conceptually, this is quite close to the first departure. The outlying value(s) will give rise to some excessively large W values, but unlike the general heavy-tailed situation, will tend not to depress the main portion of the plot. Thus outliers will be diagnosed by a large frequency in the top cell, but with overall good linearity in the plot. Thus K will tend to be near zero, while V is large and Q is small.

Some follow-up detective work to support this diagnosis will consist of checking which data points contribute to the count in the top cell.

(3) A wrong value for μ or σ. If μ is wrongly specified, then the W values will have a non-central chi-squared distribution and tend not to have enough values very close to zero. Thus if the P-P plot goes nearly horizontal near the origin but is close to linear away from the origin, a possible cause is a wrongly specified mean value μ. The standard deviation σ scales all the W values. If it is wrongly specified then all the W(x,h) values will tend to be either too small or too big, and this will lead to either a convex (σ too small) of concave (σ too big) plot. In either event, K will be large.

The shape of the plot obtained from a wrong σ is quite like that obtained when the distribution is heavier- or lighter-tailed than the Gaussian, but can be distinguished from the latter by the frequency in the top cell. With a heavy-tailed distribution, the top cell will tend to have a large frequency and a large positive V, but with a mildly underspecified σ, this tendency will not be so marked.

(4) A wrong variogram model To distinguish a wrong variogram from a wrong variance, we will have to assume that the variogram model is wrong in that τ(h) is larger than specified in some parts of the range, and smaller than specified in other parts of the range. When this is the case, at lags where the model underspecifies the variogram the W(x,h) will be too large, while where the model overspecifies the variogram the W(x,h) will be too small. The overall P-P plot and the statistics K, V and Q may be quite unremarkable here.

To deal with and diagnose this departure, it is necessary to separate out the W(x,h) terms from different lags. This leads to:-

> Proposal 2. Separate the W(x,h) into two or more subgroups according to the separations h on whi are based. Make a separate analysis of the W(x,h) in each of these subgroups, forming the plot anc finding the top cell count and the largest value. The statistics K, V and Q of each subgroup should correspond to a straight line through (0,0) and (1,1); if any group deviates from this straight line, tł variogram is misspecified.

Illustration

We now illustrate these proposals with a simulated data set and three modifications of that data set to illustrate different model departures. In all, when separating out the data by lag, we regard lags of $<\sqrt{12}$ as 'short' and those of $>\sqrt{12}$ as 'long'. The test statistics of the data sets are shown in Table 2.

(i) Clean data, good variogram model. This is a data set of values on a regular 15×15 grid defined by a normally-distributed spatial moving average process. The data are shown as Table 1. The sample variogram was found, and a spherical model fitted by ordinary least squares over the shorter lags. The sample variogram and the fitted spherical model are shown in Figure 1.

This data set has n=225 values, and so there are $\frac{1}{2}\times225\times224 = 25200$ pairwise distances W(x,h). Transforming these to the U(x,h) gives the P-P plot of Figure 2a. This plot appears to be very close to the target line, and indeed the Kolmogorov statistic of 0.0230 does not indicate any serious departure from the target line, well below our suggested guideline of $1.22/\sqrt{225} = 0.08$. The frequency of U values in the range 0.99 to 1 is 0.01254, which appears to be close to the nominal 0.01. The test statistic $V = \sqrt{n}(10P-0.1)$ is 0.381, which is far from statistically significant.

The largest W(x,h) is 19.351, for a Q value of $25200\times e^{-\frac{1}{2}\times19.351} = 4.5$, well above any reasonable significance testing threshold.

Repeating these diagnostic checks separately for short lags and for long lags similarly gives no statistical significance on any of the three tests, or any visual departure of the P-P plots (Figures 2b and 2c) from the target.

The data set thus passes the diagnostic checks proposed.

(ii) Clean data, mismodeled variogram Next we took the same data, but distorted the variogram model fitted, leaving the sill and range intact but reducing the nugget effect to zero by proportionally scaling down the fitted variogram between lag zero and the range. This calculation was intended to simulate the situation in which the variogram is fitted paying close attention to the large lags, so that a poor job is done of modeling the behavior at short lags.

The P-P plots for this analysis are shown as Figure 3 - 3a for all lags, 3b for short lags and 3c for long lags. We see that while the full data set, and the set of long lags pass the Kolmogorov screen, that at short range does not, but shows a significant tendency for a shortage of W(x,h) in the small to intermediate range of values (the plot lies above the target line).

Testing the frequency of the top class, the V statistics are 2.11, 7.91 and 1.25 respectively. This shows that the top class has about the right frequency at the long lags, but that the frequency is much too high at the short lags and that this also somewhat over-represents the top class in the full set of distances.

The Q statistics are 2.1×10^{-6} for the full data set, 3.2×10^{-7} for the short lags, but 0.6 for the long lags. This again indicates a good model fit for long lags, but poor for short lags.

<u>(iii) Non-normal data, good variogram model</u> To see the effect of heavy-tailed non-normality, we created a third data set by transformation from the first. If Z represents a value in the original data set, then the transform was to

$$Z^* = e^{(1+0.25Z)}.$$

The resulting log-normally distributed values, having logarithmic mean 3.55 and logarithmic standard deviation 0.57 were heavier-tailed than normal, but not very extreme.

The sample variogram and fitted spherical model are shown as Figure 4. The largest W(x,h) value is 36.307, corresponding to $Q(x,h) = 1.3\times10^{-8}$, a figure far too small to plausibly arise by chance in a data set conforming to the model. For the short and long lags also, the Q values are very small.

Figure 5 shows the P-P plots - 5a for the full data set, 5b for short lags and 5c for long lags. All three clearly show the convex shape associated with heavy-tailed data, and give Kolmogorov statistics - of 0.108; 0.114 and 0.107 well above the threshold of 0.08.

The V statistics are 4.17, 4.23 and 4.16 respectively, showing an excess at both short and long lags of very large W(x,h) values.

Putting these three sets of figures together, several things are clear - the model does not fit, but the problem does not seem to be lag specific. This suggests that either the overall variance was badly misspecified or the data are non-normal. It is an easy calculation to verify that the former is not the problem.

(iv) Outlying values. Finally we introduced two bad values into the data, replacing the value 12.297 in the (5,7) position of the data and the value in the (11,3) position with 2.590 (the smallest of the good values seen). Neither of these introduced outliers is detectable by univariate methods, since both values are within the range of values of the good data, and their outlyingness can be detected only by their aberration from the values of their neighbors.

The largest W(x,h) value for the altered data set is 26.86, for a Q value of 0.0371, which demonstrates outlyingness at a conservative 5% significance level. This outlyingness is verified in the short and long lags groups separately.

The P-P plots are shown as Figure 6a (all lags), 6b (short lags) and 6c (long lags) respectively. None of the plots shows any noted departures from the target line, and the Kolmogorov statistics of 0.038 0.029 and 0.038 are well below the threshold of 0.08, indicating no systematic departures from the model specified.

The frequencies in the top class however are 0.031 (all data), and 0.033 (short lags) and 0.030 (long lags). These correspond to standard normal values of V = 3.08, 3.55 and 3.02 respectively. These values are highly significant, showing an excess of very large W(x,h) values.

The overall conclusions then is of a good linearity of the P-P plot at all lags, but an excess of values in the high extremes. This correctly diagnoses the presence of outliers in an otherwise good model.

Discussion

There are many procedures for testing a sample of independent values for normality, but one of the most attractive is the normal probability plot. This plot provides an immediate visual assessment not only of the extent to which the data conform to normal, but also, in the event that the data do not, of the nature of the violation.

When one goes from an independent sample to a regionalized variable however, the situation is considerably complicated. Not only does one now wish to test for the individual normality of the data, but of multivariate normality, and also of concordance to the variogram specified for the data.

Diamond and Armstrong (1984) point out the potential severe impact on the kriging weights of apparently minor changes in the fitted variogram model. However if two different models give vastly different kriging weights, but one model fits the data while the other does not, then the sensitivity of the weights to the model is of academic concern only. This points out the need for ways of assessing whether a particular data set does or does not conform to one or more specified variogram models.

Checking for normality has slightly different objectives. When one is performing linear kriging, then multivariate normality of the data implies that linear kriging is optimal and has strong lack of bias properties, a situation whose pleasant consequences one would like to verify when it pertains. More importantly however if lognormal kriging is being considered, then the question of the lognormality of the data becomes more crucial. Only if the data really truly lognormal are the standard methods of inverting the log transformation valid. Thus a test of the multivariate normality after log transformation fills a very basic need of lognormal kriging - see for example Rivoirard (1990).

In this paper, we propose a procedure which tests the normality and the variogram simultaneously. The test is of the 'necessary but not sufficient' type. It is possible for data to be bivariate normal but not multivariate normal (a situation which the procedure cannot detect). It is also possible for the quadratic form tested to have a chi-squared distribution even if the two underlying variables do not have a bivariate normal distribution. We believe however that these two possible cracks through which the procedure can fall are not very wide in real data sets. Thus if the data set fails the test it is certain that the normality, the modeled variogram or both are wrong, but if the data set passes the test it is likely though not certain that the normality and assumed variogram model are correct.

References

Diamond, P., and Armstrong, M., (1984), 'The robustness of variograms', Mathematical Geology, 16, 809-822.

Cox, D. R., and Small, N. J. H., (1978), 'Testing multivariate normality', Biometrika, 65, 263-272.

Hawkins, D. M., (1981), 'A new test for normality and homoscedasticity', Technometrics, 23, 105-110.

Rivoirard, J., (1990), 'A review of lognormal estimators for in situ reserves', Mathematical Geology, 22, 213-221.

Table 1. Input data for illustrations

Row															
1	9.615	12.754	9.092	9.266	8.928	7.997	8.718	5.445	10.488	9.457	12.277	13.895	7.863	8.068	4.539
2	7.757	7.017	9.921	9.802	8.616	10.160	12.042	13.778	12.291	11.837	13.027	13.354	10.577	6.428	4.444
3	7.213	6.926	8.303	8.285	8.298	11.698	11.925	10.459	9.931	13.060	12.112	11.553	8.280	2.590	4.585
4	9.327	9.602	7.555	8.073	11.335	12.370	9.568	11.038	11.611	9.768	9.189	6.474	9.634	6.633	5.795
5	9.946	7.777	9.192	11.461	13.375	11.803	12.297	11.287	7.035	10.072	8.228	8.804	7.786	8.598	8.824
6	11.712	8.698	11.559	12.492	12.744	14.286	11.831	10.546	6.154	9.787	10.339	10.509	8.602	5.650	9.668
7	12.111	7.830	10.051	10.998	10.195	12.266	12.300	8.034	9.463	8.108	7.690	5.724	8.011	8.195	5.020
8	11.225	9.368	10.000	11.299	11.116	13.162	8.577	8.811	10.308	11.063	7.532	7.946	7.545	8.605	5.745
9	9.561	8.624	8.671	10.276	11.587	13.956	8.729	8.874	13.017	10.556	9.482	10.303	10.177	9.555	11.099
10	11.569	9.276	10.381	12.404	12.311	10.070	9.758	11.488	11.353	15.248	9.874	11.815	7.239	7.000	9.775
11	12.204	10.727	12.071	12.893	10.474	13.163	8.153	11.612	10.145	13.589	13.292	11.323	10.698	12.677	12.435
12	8.977	11.514	12.476	14.983	14.014	8.686	9.050	10.937	12.632	14.220	12.112	9.206	11.952	13.035	9.372
13	8.302	10.016	11.023	12.504	9.367	6.727	10.751	11.093	12.270	14.705	10.507	5.941	7.380	11.700	8.132
14	10.446	10.881	14.780	13.529	11.294	9.834	11.128	11.380	10.522	11.255	12.090	8.619	12.301	13.772	10.293
15	9.474	9.902	11.415	11.236	10.240	12.423	11.251	11.471	10.644	12.800	16.096	10.855	10.712	11.822	11.865
Col	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

Table 2. Test statistics

Data set	Statistics	K	V	Q
	Cutoff values	0.08	± 2	0.05
Good data	all lags	0.022	0.33	4.5
	short lags	0.034	0.18	0.6
	long lags	0.020	0.35	6.4
Wrong γ	all lags	0.045	2.11	2.1×10^{-6}
	short lags	0.125	7.91	2.7×10^{-7}
	long lags	0.033	1.25	0.60
Heavy tails	all lags	0.108	4.17	4.1×10^{-4}
	short lags	0.114	4.23	1.7×10^{-3}
	long lags	0.107	4.16	3.6×10^{-1}
Outliers	all lags	0.038	3.09	0.037
	short lags	0.029	3.55	0.057
	long lags	0.039	3.02	0.032

Figure 1.

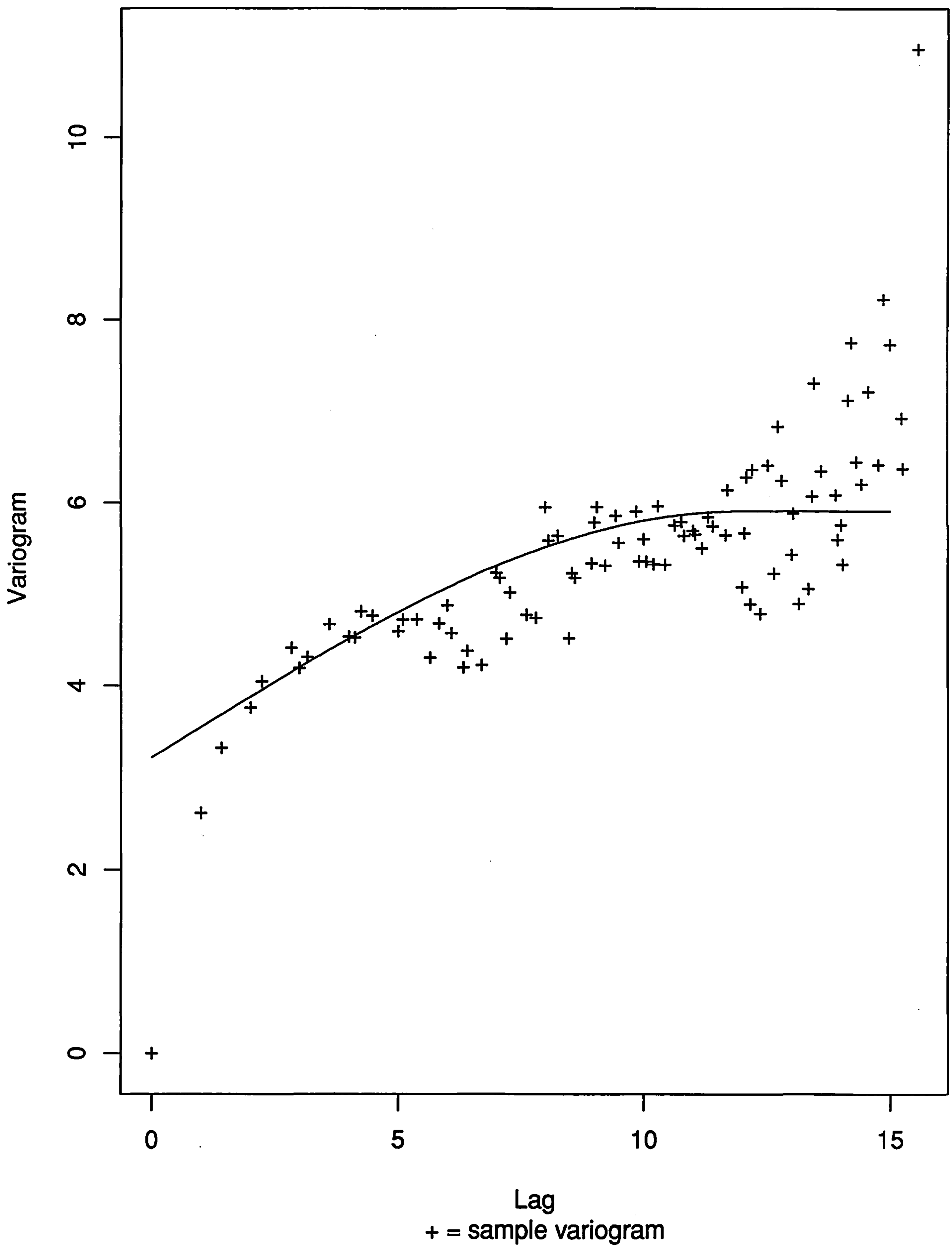

Figure 2a.

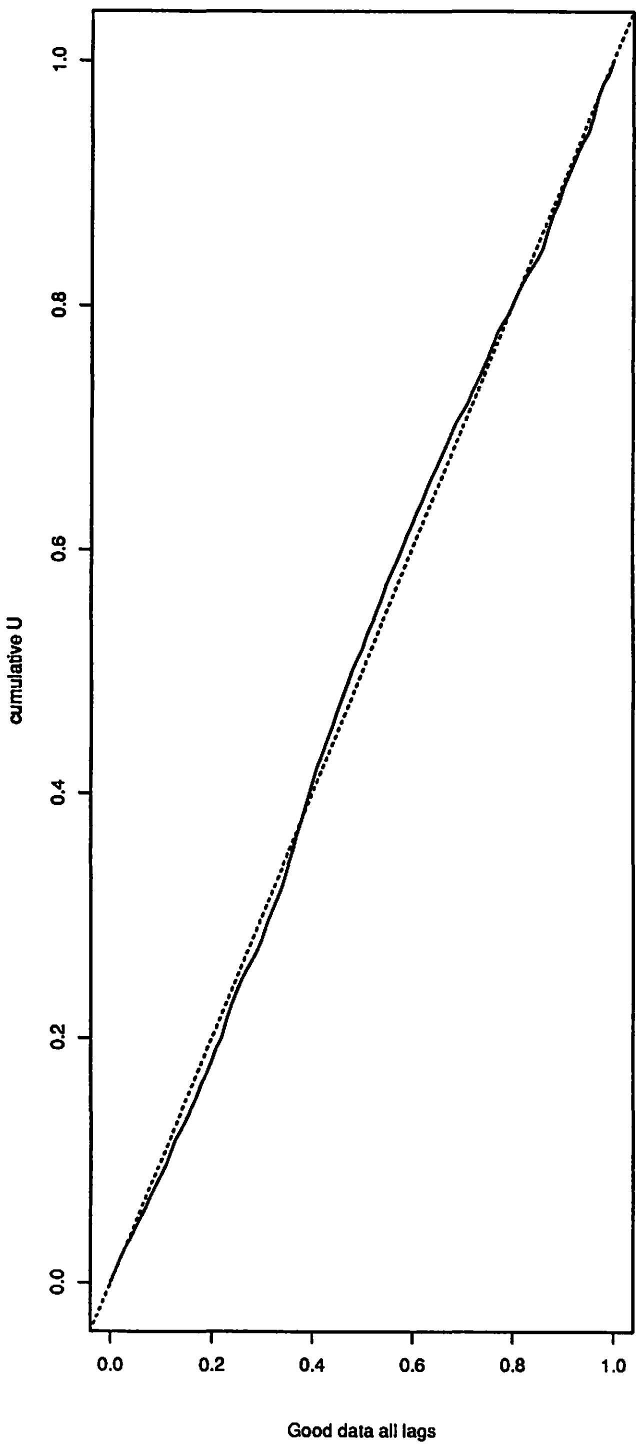

Figure 2b.

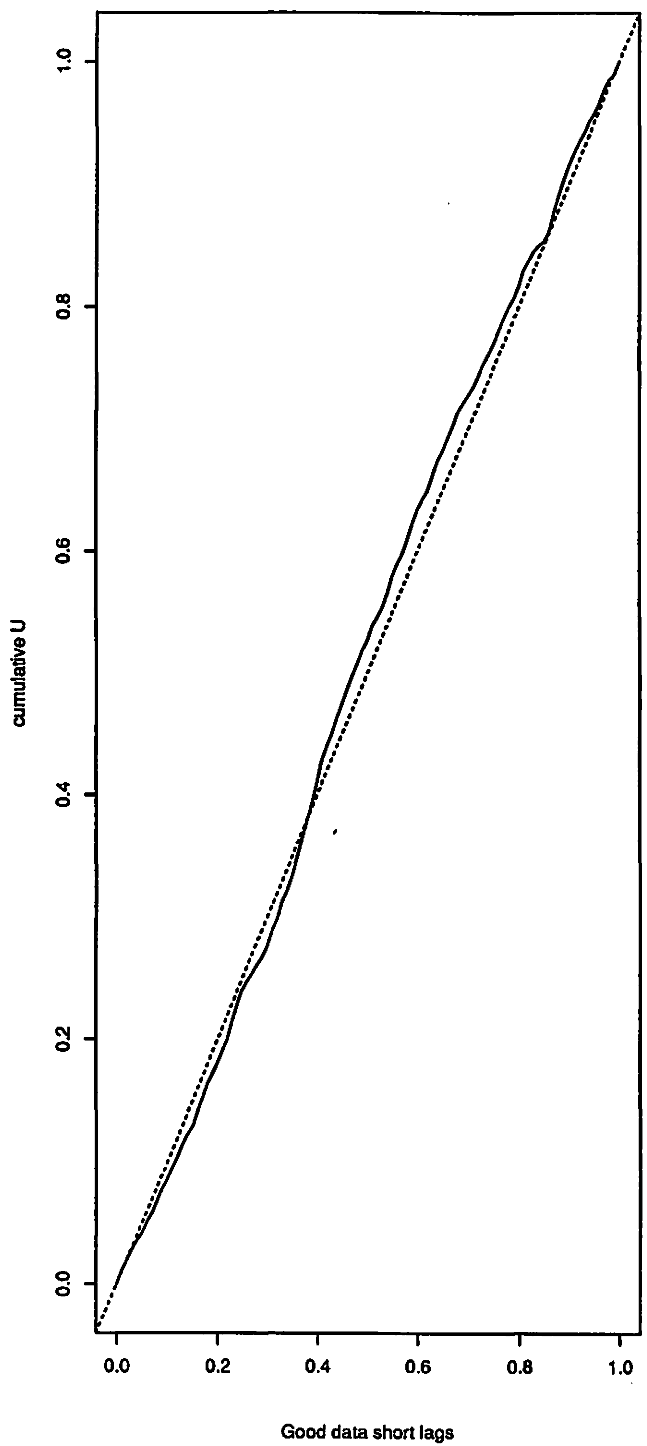

Figure 2c.

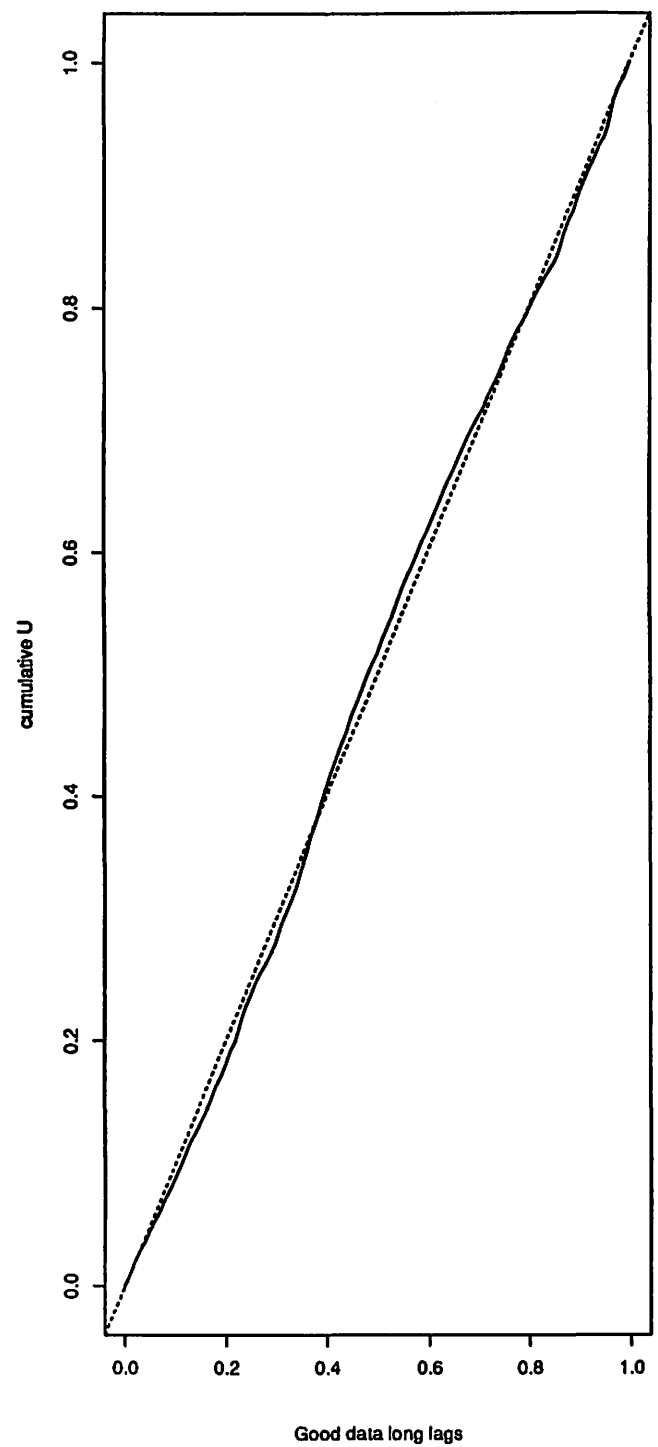

Figure 3a.

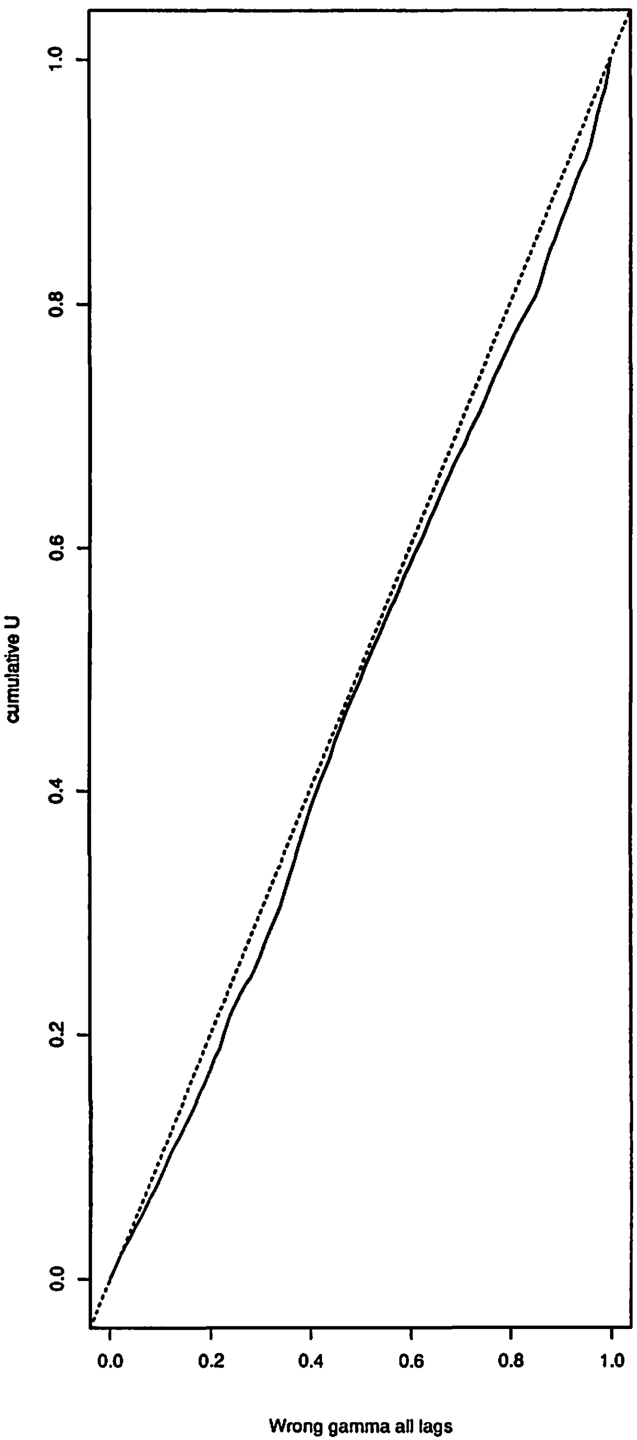

Figure 3b.

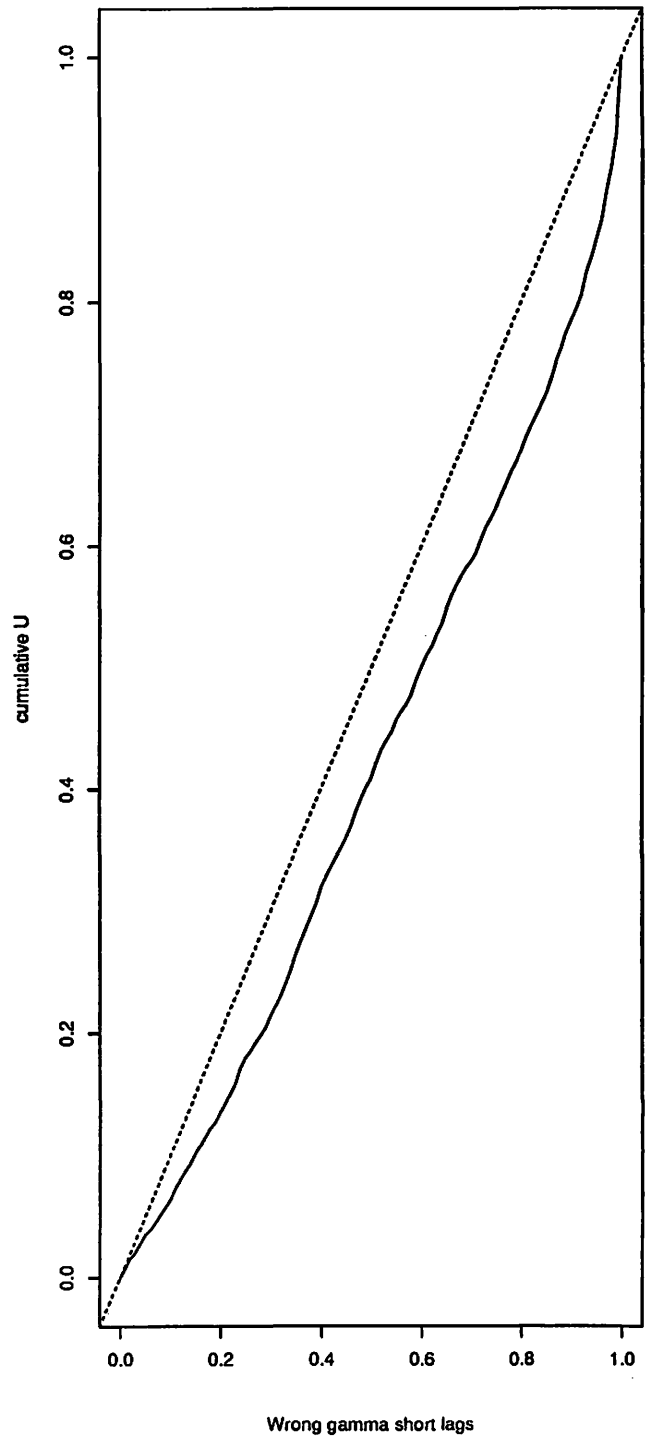

Figure 3c.

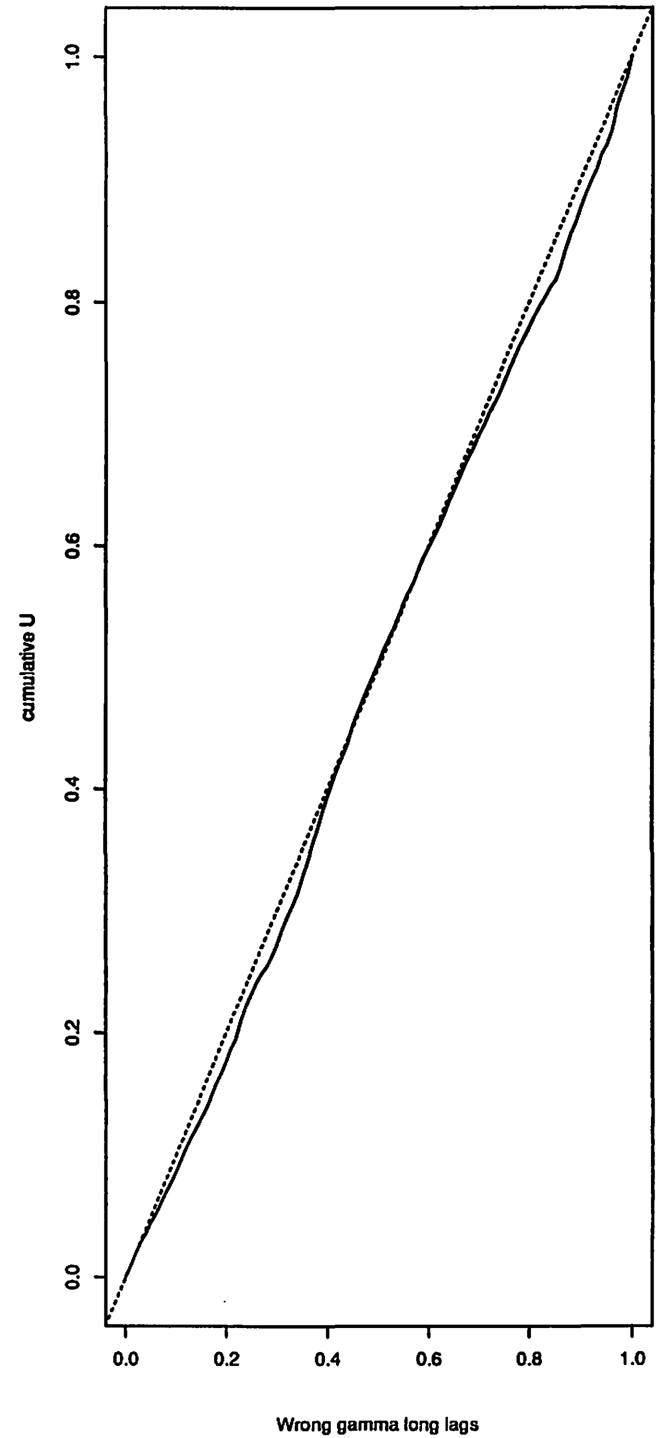

Figure 4.

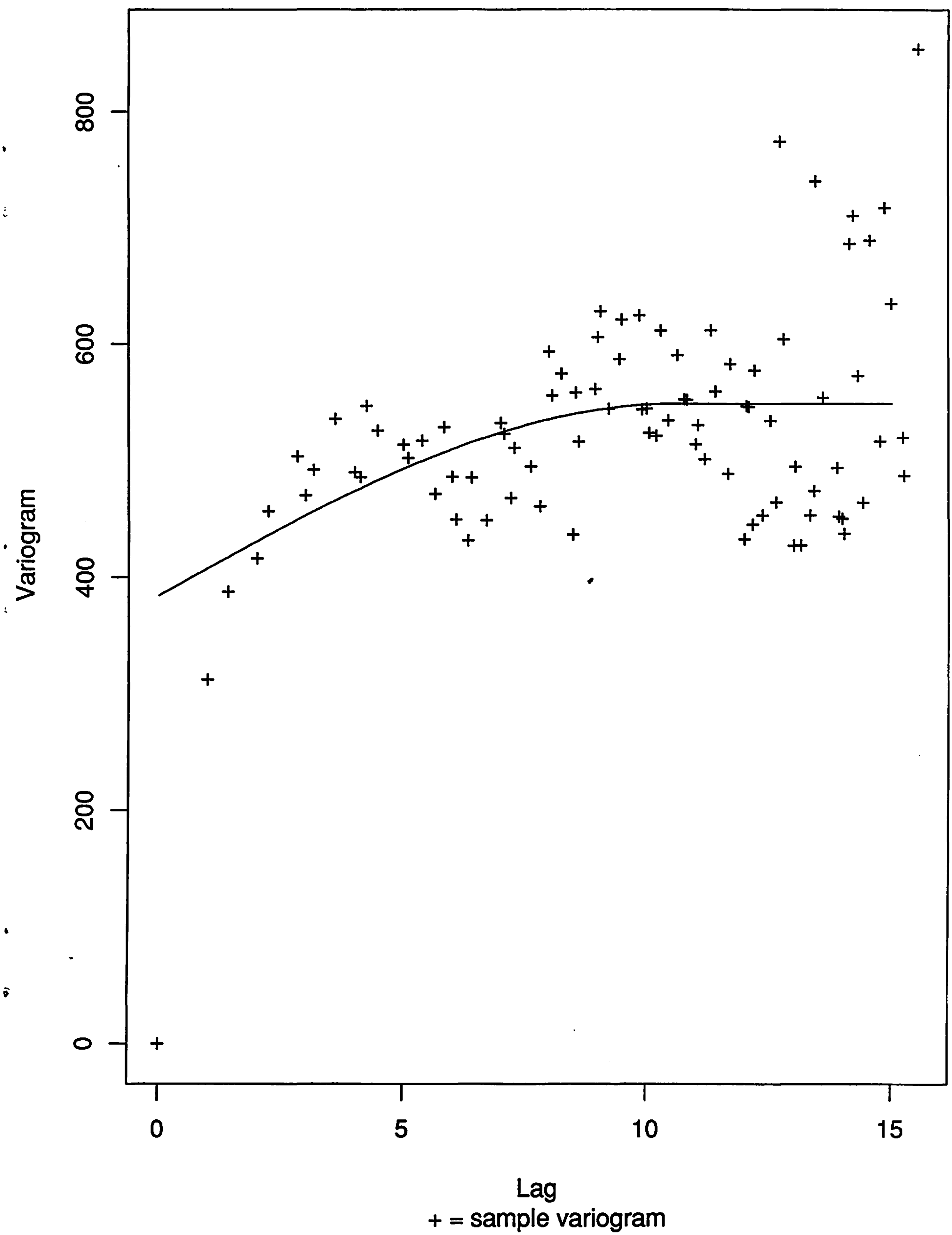

Figure 5a.

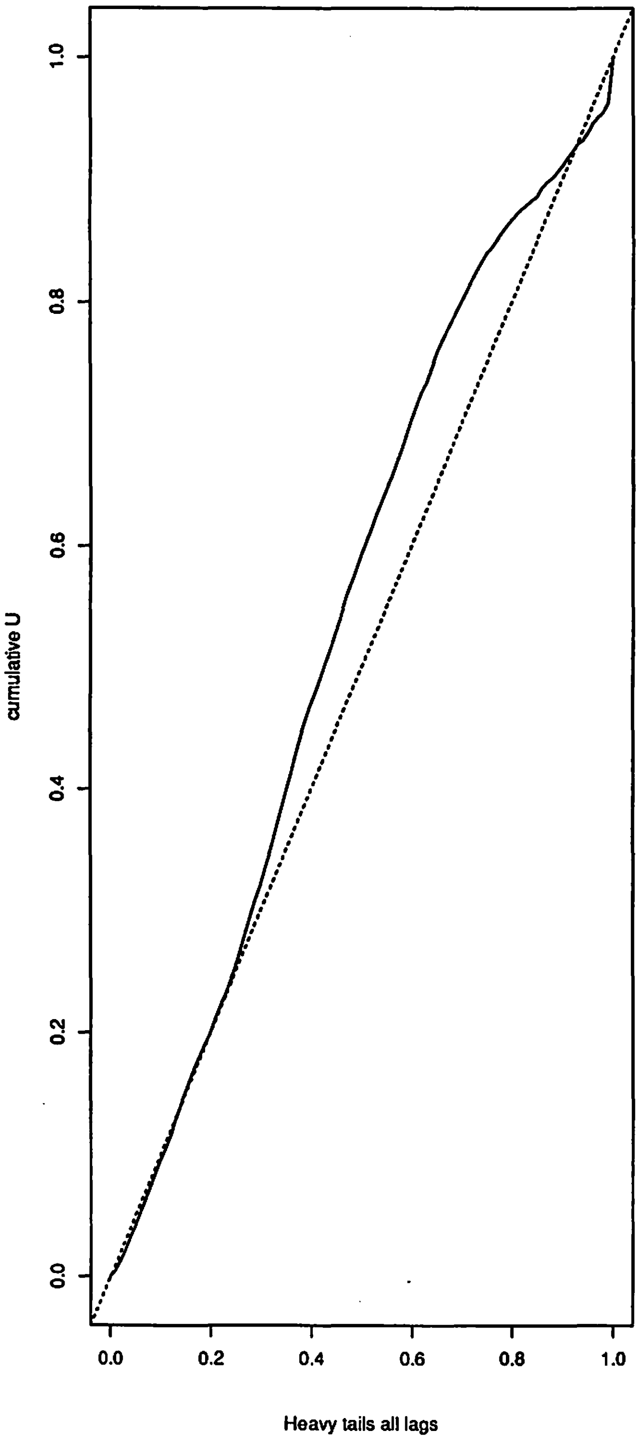

Figure 5b.

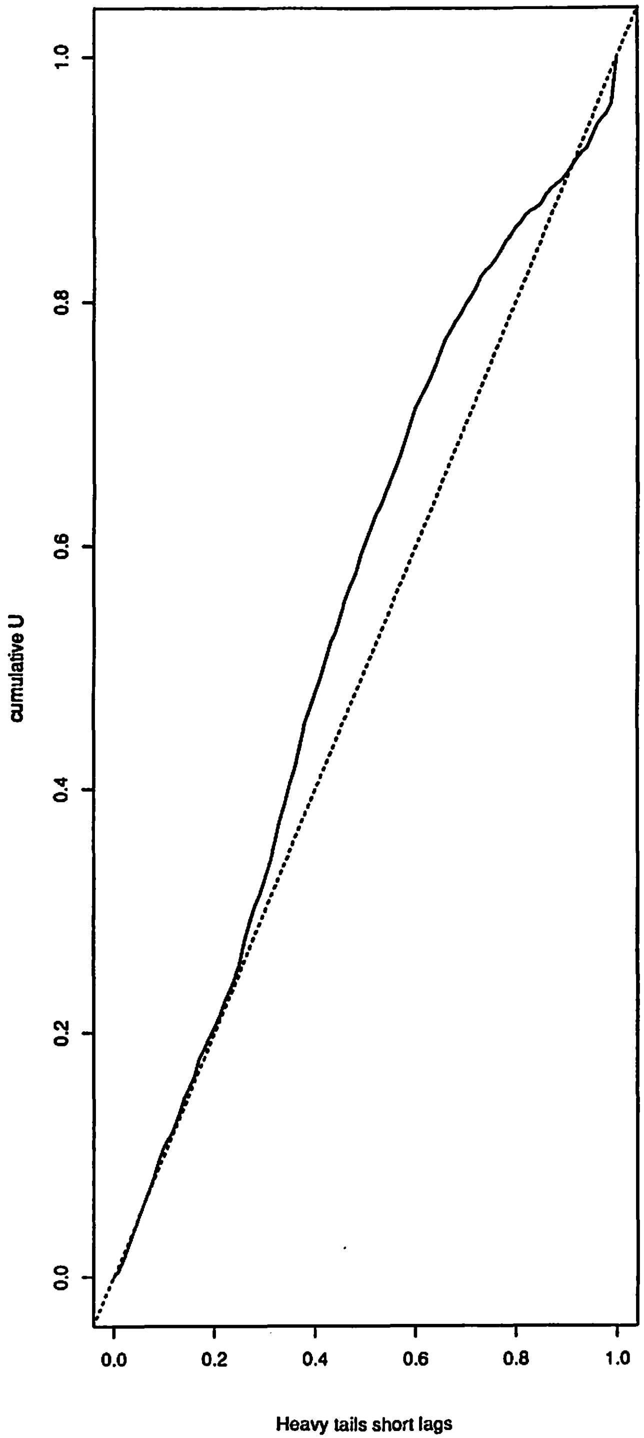

Figure 5c.

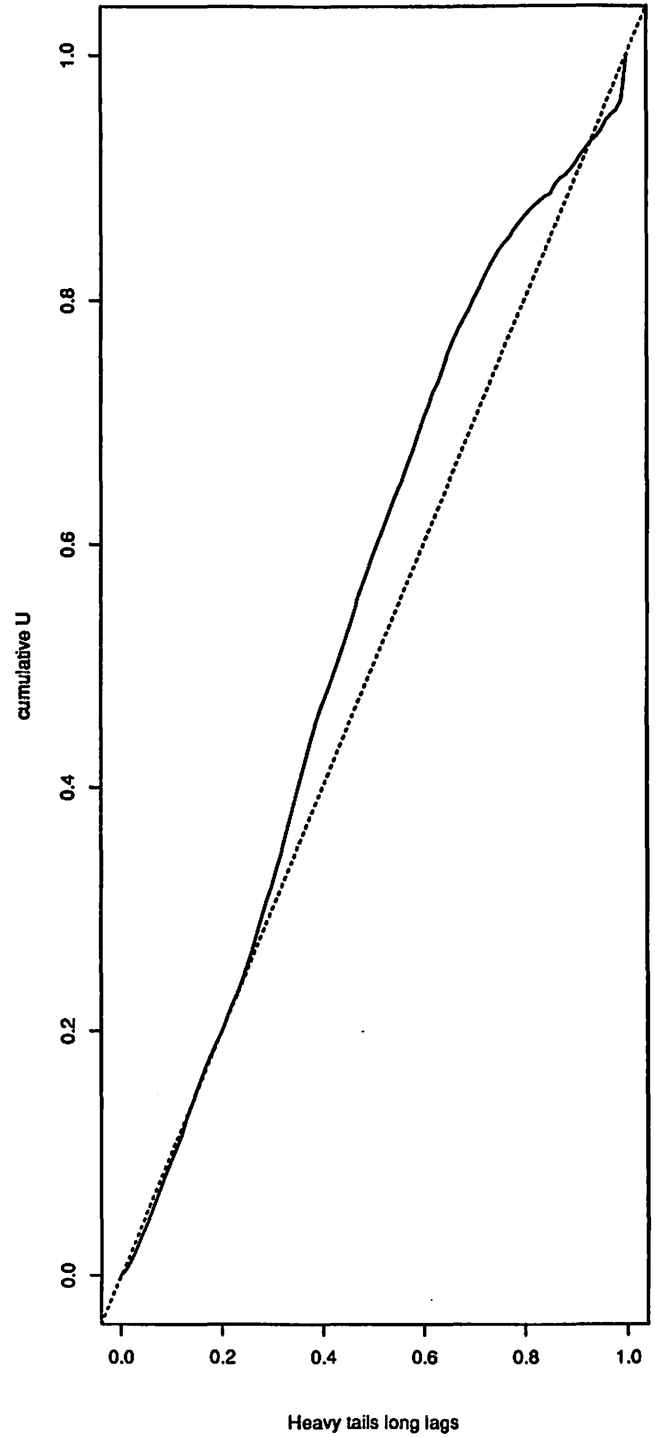

Figure 6a.

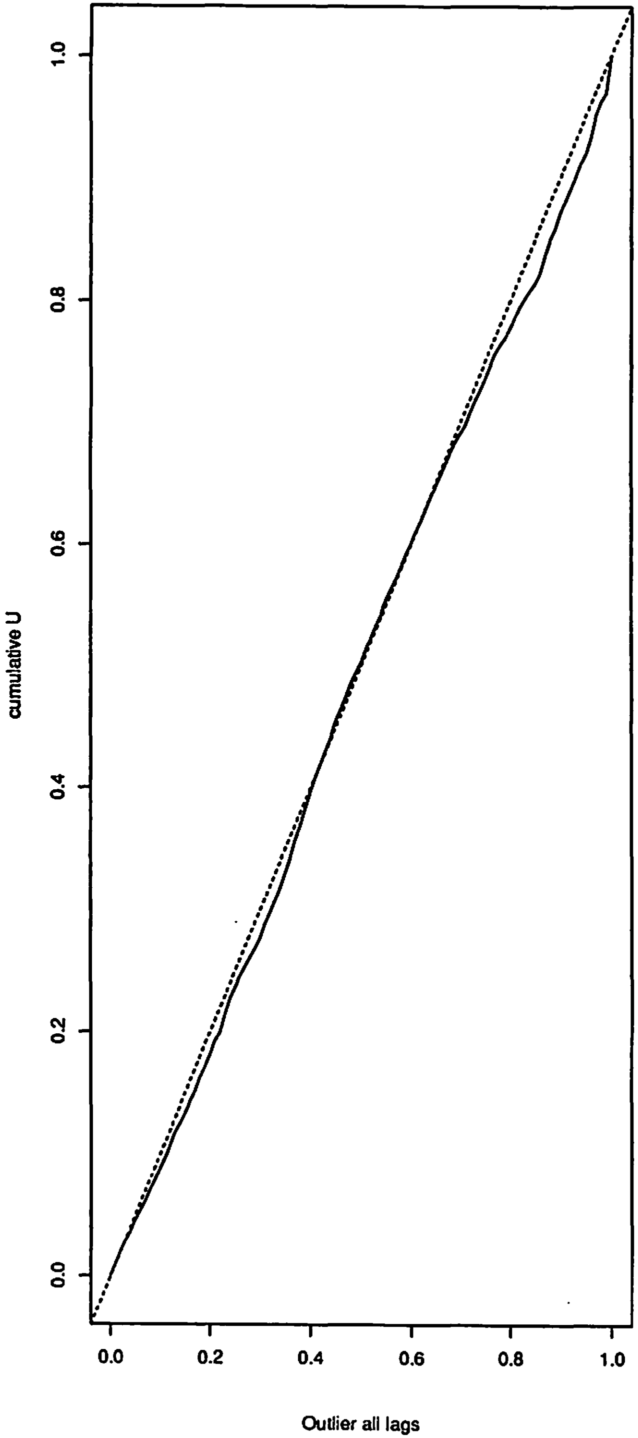

Figure 6b.

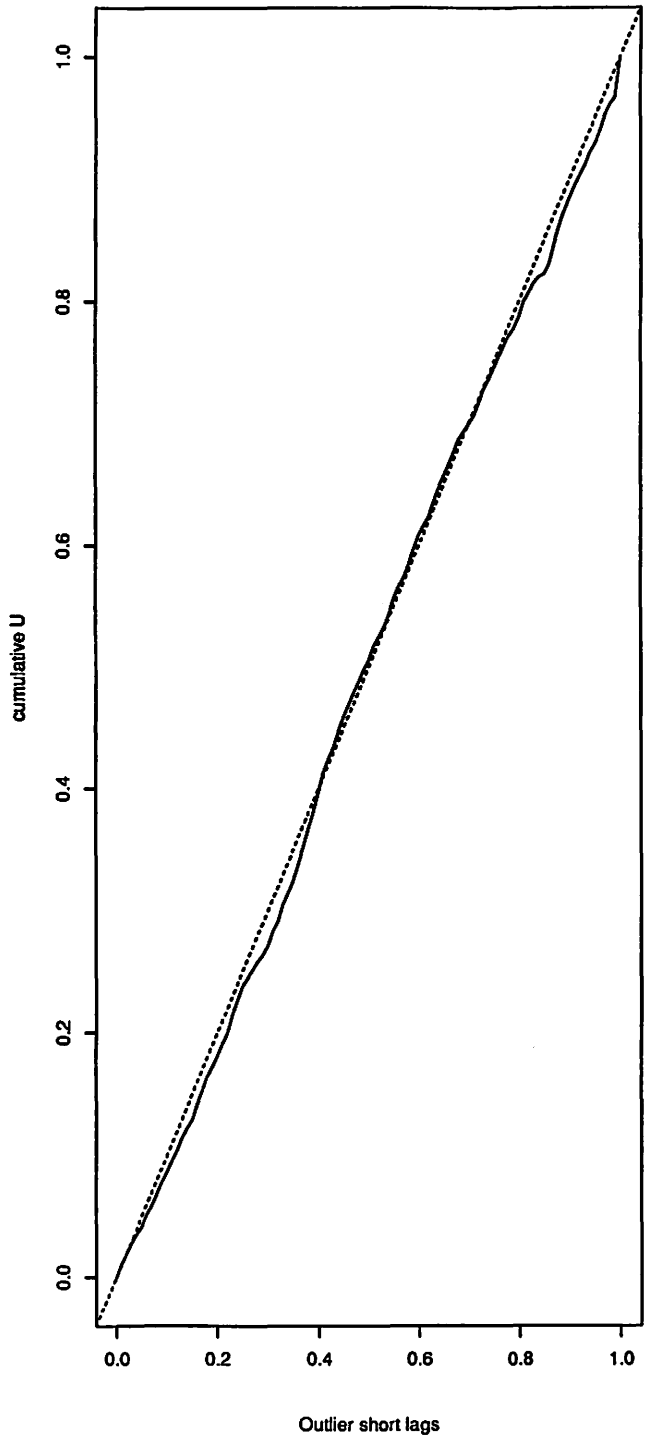

Figure 6c.